Mellanox Virtual Modular Switch™ Reference Guide

Rev 1.0

www.mellanox.com

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCTO(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies, Ltd.
Beit Mellanox
PO Box 586 Yokneam 20692
Israel
www.mellanox.com
Tel: +972 (0)74 723 7200
Fax: +972 (0)4 959 3245

© Copyright 2013. Mellanox Technologies. All Rights Reserved.

Mellanox®, Mellanox logo, BridgeX®, ConnectX®, CORE-Direct®, InfiniBridge®, InfiniHost®, InfiniScale®, MLNX-OS®, PhyX®, SwitchX®, UFM®, Virtual Protocol Interconnect® and Voltaire® are registered trademarks of Mellanox Technologies, Ltd.

Connect-IB™, FabricIT™, Mellanox Open Ethernet™, Mellanox Virtual Modular Switch™, MetroX™, MetroDX™, ScalableHPC™, Unbreakable-Link™ are trademarks of Mellanox Technologies, Ltd.

All other trademarks are property of their respective owners.

Document Revision History

Table 1 - Document Revision History

Document Revision	Date	Description
Rev. 1.0	July 2013	Initial release.

Mellanox® Technologies Confidential

About this Document

This reference architecture provides general information concerning Mellanox Virtual Modular Switch configuration and design.

Intended Audience

This manual is intended for field engineers and network administrators intending to design a data center topology using Mellanox Ethernet switch platforms.

Related Documentation

The following table lists the documents referenced in this document.

Table 2 - Reference Documents

Document Name	Description
MLNX-OS® User Manual	This document contains feature description and configuration flows. The document can be retrieved in Mellanox support page.
MLNX-OS® Command Reference Guide	Command Reference Guide for MLNX-OS listing all of the commands available through MLNX-OS with explanations and examples. The document can be retrieved in Mellanox support page.
Mellanox Ethernet Switch Systems	Mellanox SX1016, SX1024, SX1036 The User Manual and Product brief can be retrieved here: http://www.mellanox.com/content/pages.php?pg=ethernet_switch_overview&menu_section=71
Mellanox Approved Cable List	Mellanox List of Approved Cables. The document can be retrieved here: http://www.mellanox.com/related-docs/user_manuals/Mellanox_approved_cables.pdf

Mellanox® Technologies Confidential

Document Revision History 3
About this Document 4
Intended Audience 4
Related Documentation 4
Chapter 1 Introduction 6
1.1 Design Considerations 6
1.2 Routing 6
Chapter 2 VMS Architecture 8
2.1 Mellanox® VMS-72 (Entry Level) 8
2.2 Scaling 8
2.3 Mellanox VMS-648 (Full Size) 9
2.4 ToR Connectivity 10
Chapter 3 VMS Configuration 11
3.1 Prerequisites 11
3.2 Common Procedures 11
3.2.1 Creating a VLAN Interface 11
3.2.2 Creating a Static LAG 12
3.2.3 Setting a VLAN Association to L2 interface 13
3.2.4 Basic OSPF Configuration 13
3.3 Configuration Example 14
3.3.1 Spine-1 15
3.3.2 Spine-2 16
3.3.3 Leaf-1 17
3.3.4 Leaf-2 17
3.3.5 Leaf-3 18
3.3.6 Leaf-4 18
Chapter 4 Advanced Considerations 19
4.1 Load Balancing 19
4.1.1 Hash Functions 19
4.1.2 Equal Cost Multi-Path Routing (ECMP) 20
4.2 56GbE Interface Rate 21
4.3 Store and Forward 21
4.4 OSPF 21
4.4.1 Totally Stub Area OSPF 21
4.5 Automation Tools 25
4.6 Network Discovery 25
Appendix A Ordering Information 27
Appendix B Mellanox VMS – Basic POC 28

Mellanox® Technologies Confidential

1 Introduction

As new applications are constantly evolving, data centers must be flexible and future-proof to meet the demand for higher throughput and better scalability, while protecting the investment, without increasing the power consumption or cost.

Traditionally, the aggregation switching of data centers of cloud providers, Web 2.0 providers and enterprises were based on modular switches which are usually expensive to purchase and operate, and are optimized for specific cluster sizes. These modular switches do not provide the required flexibility and tend to lag behind the progress of data center technologies.

To overcome this limitation of the modular switches, more fixed switches (or top-of-rack switches) are used to increase efficiency in data center aggregation.

The Mellanox Virtual Modular Switch™ solution (VMS), comprised of Mellanox® SX1036 switch systems, provides an ideal optimized approach for new levels of flexibility, scalability and efficiency, while future-proofing the investment and reducing expenses.

This document presents Mellanox VMS topology and configuration and offers a set of guidelines for advanced users. Naturally, network designers may find it necessary to apply some topology modifications to better suit their needs.

1.1 Design Considerations

Each data center has its own characteristics and needs. In general, when designing a large data center network, it is critical to consider the following variables:

- Interface rate – server clusters may be comprised of tens of thousands of ports, thus requiring high-performance packet switching, large bandwidth, and high port speeds. Server interfaces are moving away from 1GbE to 10GbE and 40GbE. Beyond the standard 10GbE and 40GbE link rates, Mellanox systems also offer a 56GbE proprietary speed that reduces costs and rack space in the data center.
- Number of servers – the number of servers (overall and per rack) multiplied by the interface rate gives the north-bound throughput of the ToR. The number of servers per rack normally reaches 40 servers.
- Oversubscription ratio – data center applications may require different oversubscription ratios. North-South traffic and East-West traffic have different characteristics according to the network application. Also, non-blocking or any blocking ratio is possible. The oversubscription ratio must be examined at multiple aggregation points of East-West and North-South traffic.
- Latency – latency affects the overall performance of the cluster. Using Mellanox low-latency switch platforms helps ensure optimal performance. However, low-latency may not always be a critical factor in the cluster design. Certain applications might prioritize bandwidth over latency, which might impact topology decisions when designing the network.

1.2 Routing

The IP network topology is discovered via the Open Shortest Path First (OSPF) routing protocol that is enabled on all ToR and aggregation devices in the network. OSPF routers exchange link status messages with their physical neighbors in order to establish the same network topology for all routers in the network.

Mellanox® Technologies Confidential

The mechanism enabling the selection of several routes (dynamic routing) to the same destination is called ECMP (equal cost multi-path). Without ECMP, the router selects the best equal cost route and uses it as the only path to the destination IP address. ECMP can also be implemented over static routes as long as all routes to the same destination are indeed of equal cost.

The advantage of running ECMP with a routing protocol such as OSPF is that it supplies good and redundant load balancing. OSPF handles the dynamic learning of the network and the active routes are pushed into the routing table, while the ECMP does the load balancing between the routes.

Mellanox® Technologies Confidential

2 VMS Architecture

The VMS principal topology is a CLOS-3 that consists of a fat tree (multi-root tree) of routing elements (leafs and spines). Each of the leafs needs to physically connect to each of the spines in order to guarantee exactly two hops between any two leafs. The multiple routes between the leafs have the same weight and receive the same priority in each of the routing tables.

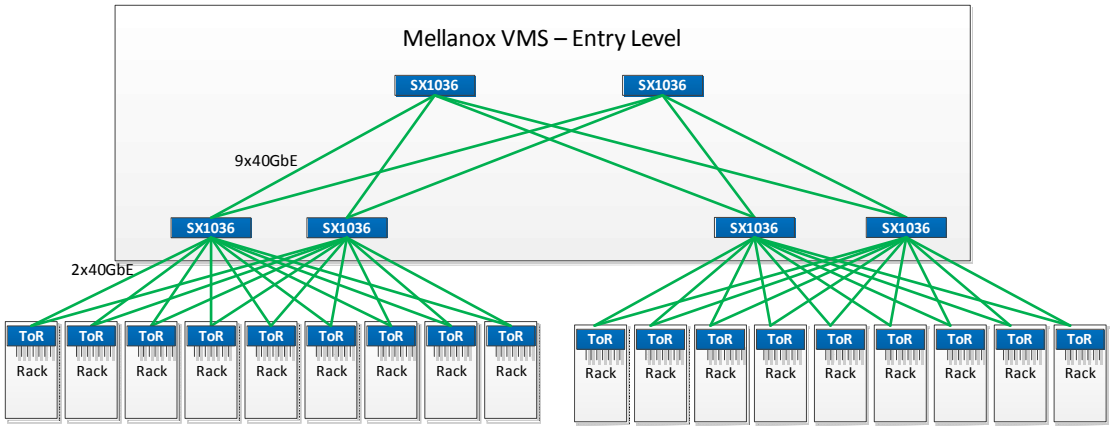
The topologies lean on non-blocking architecture design between the leafs and spines of the VMS.

2.1 Mellanox® VMS-72 (Entry Level)

Mellanox's VMS-72 (72 ports) is illustrated in the following diagram. It is a typical entry level VMS topology. It can be used for small-scale clusters or as a starting point for large scale clusters.

Mellanox's VMS-72 is equipped with 2 spines and 4 leafs. In addition, up to 18 ToR switches can be connected to the VMS assuming each ToR has 4x40GbE uplink ports configured as 2x40GbE LAG for each leaf.

Figure 1: Entry Level VMS



2.2 Scaling

The VMS scales easily by adding spines and leaf switches. The minimal number of switches required for each expansion are one spine and two leafs that add 36 extra ports to the VMS.

Best-practice VMS scaling sizes are described in Table 3. Other VMS designs are possible but must be thought through carefully as it may cause asymmetry within the VMS.

Table 3 - Best Practice of Mellanox VMS Size

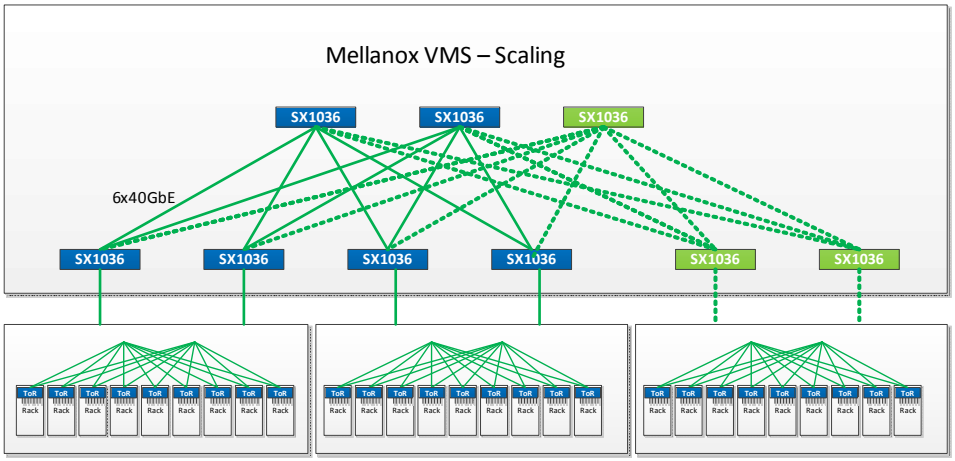
Mellanox VMS	Number of Spines	Number of Leafs	Link Size between a Leaf and a Spine
Mellanox VMS-72	2	4	9x40GbE
Mellanox VMS-108	3	6	6x40GbE
Mellanox VMS-216	6	12	3x40GbE

Mellanox® Technologies Confidential

Table 3 - Best Practice of Mellanox VMS Size

Mellanox VMS	Number of Spines	Number of Leafs	Link Size between a Leaf and a Spine
Mellanox VMS-324	9	18	2x40GbE
Mellanox VMS-648	18	36	1x40GbE

Figure 2: VMS Scaling

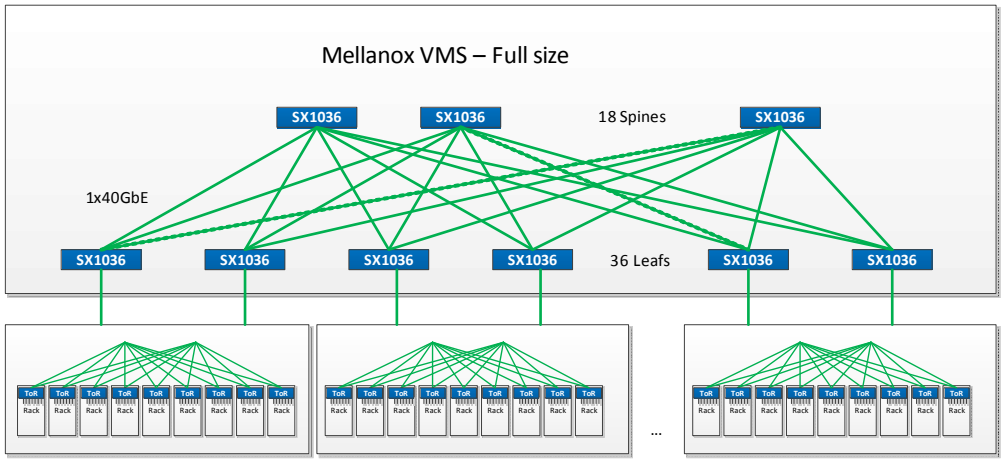


2.3 Mellanox VMS-648 (Full Size)

Mellanox VMS-648 (648 ports) is illustrated in the following diagram. It is a typical full-scale VMS topology. It is meant to be used for large clusters.

Mellanox VMS-648 is equipped with 18 spines and 36 leafs. In addition, up to 162 ToR switches (9 ToR switches * 18 leaf pairs) may be connected to the VMS assuming each ToR has 4x40GbE uplink ports configured as 2x40GbE LAG for each leaf.

Figure 3: Full Size VMS

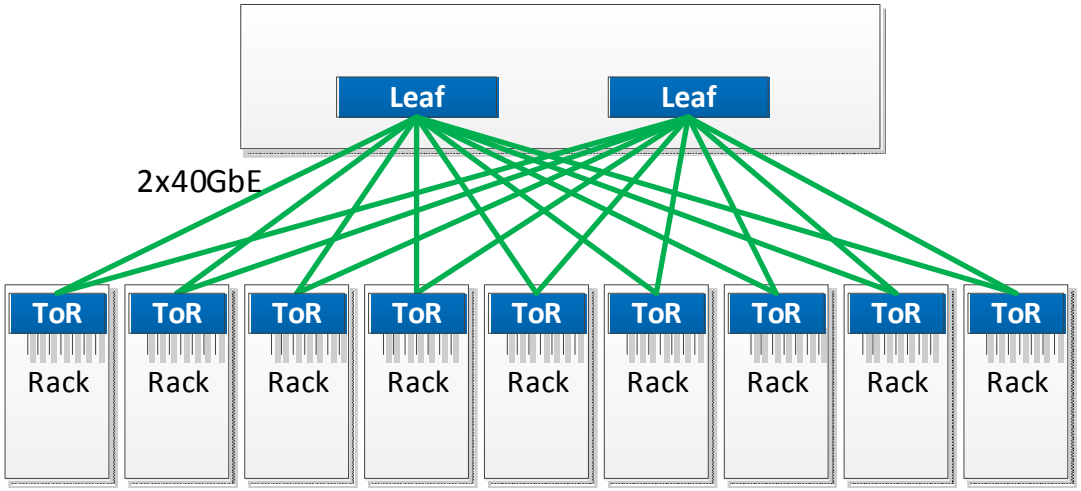


Mellanox® Technologies Confidential

2.4 ToR Connectivity

Typical ToR connectivity in the Mellanox VMS is 4x40GbE to two leafs. The link is divided to 2x40GbE per leaf switch configured as LAG. Additional configuration options are possible according to the data center's needs (for example, using a different number of links between the ToR and the leafs, or not using a high availability configuration scheme).

Figure 4: ToR Connectivity



Mellanox® Technologies Confidential

3 VMS Configuration

Once the Mellanox® VMS topology is defined network configuration must be applied to the VMS.

3.1 Prerequisites

Before attempting to configure a switch within the Mellanox VMS, make sure the following conditions are met:

- The switch system is up and has a management IP address
- The system profile is: eth-single-switch. Use the command `show system profile` to verify this.

3.2 Common Procedures

This section describes the following set of basic common procedures:

- "Creating a VLAN Interface"
- "Creating a Static LAG"
- "Setting a VLAN Association to L2 interface"
- "Basic OSPF Configuration"

3.2.1 Creating a VLAN Interface

➢ ***To create a VLAN interface (IP interface):***

Step 1. Create a VLAN using the command `vlan <vlan-id>`.

Step 2. Attach at least one interface to this VLAN. Run the command `switchport` in the interface configuration mode.

There must be at least one interface in the operational state "up".

Step 3. Create a VLAN interface using the command `interface vlan`.

Step 4. Set an IP address and subnet mask for this interface.

Example:

```
switch (config)# vlan 10
switch (config vlan 10)# exit
switch (config)# interface ethernet 1/1
switch (config interface ethernet 1/1)# switchport mode access
switch (config interface ethernet 1/1)# switchport access vlan 10
switch (config interface ethernet 1/1)# exit
switch (config)# show interface etherent status  // ensuring there is a port in UP
state
```


Mellanox® Technologies Confidential

```
Port                  Operational state          Speed                 Negotiation
----                  -----------------          -----                 -----------
Eth1/1                Up                       40 Gbps               No-Negotiation
…



switch (config)#interface vlan 10
switch (config interface vlan)#ip address 10.10.10.10 /24
switch (config interface vlan)#show interface vlan 10

Vlan 10
 Admin state: Enabled
 Operational state: UP
 Mac Address: 00:02:c9:5d:e0:f0
  Internet Address: 10.10.10.10/24
  Broadcast address: 10.10.10.255
  MTU: 1500 bytes
  Description: my-ip-interface
  Counters: disabled
switch (config interface vlan)#
```

3.2.2 Creating a Static LAG

➢ ***To create a static LAG:***

Step 1. Create a port-channel (LAG) using the command `interface port-channel`.

Step 2. Add physical interfaces to the port-channel group using the command `channel-group` in the interface configuration mode.

Optionally, LACP may be enabled.
Refer to *MLNX-OS® User Manual* (Link Aggregation Group chapter), for more information.

Example:

```
switch (config) # vlan 10 // creating a vlan (if nonexistent)
switch (config vlan 10) exit
switch (config) # interface port-channel 1
switch (config interface port-channel 1) #exit
switch (config) # interface ethernet 1/1
switch (config interface ethernet 1/1) # channel-group 1 mode on // static LAG
```

If the physical port is operationally up, this port becomes an active member of the aggregation. Consequently, it becomes able to convey traffic.

Mellanox® Technologies Confidential

3.2.3 Setting a VLAN Association to L2 interface

➢ ***To associate a VLAN to L2 interface (Ethernet or LAG):***

Step 1. Enter the interface configuration mode (Ethernet or port-channel)

Step 2. Set a VLAN ID to an interface using the command `switchport`.

Example for Ethernet Interface:

```
switch (config) # interface ethernet 1/1
switch (config interface ethernet 1/1) #switchport access vlan 10
```

Example: for LAG Interface:

```
switch (config) # interface port-channel 1
switch (config interface port-channel 1) #switchport access vlan 10
```

3.2.4 Basic OSPF Configuration

➢ ***To configure OSPF:***

Step 1. Enable OSPF configuration using the command `protocol ospf`,

Step 2. Create a router OSPF instance using the command `router ospf`.

Step 3. Associate the required IP interfaces to the OSPF area using the command `ip ospf area`.

Step 4. It is recommended that all the internal links in the VMS are configured as point-to-point OSPF links for fast convergence. Use the command `ip ospf network point-to-point` to accomplish that.

Step 5. For links connected directly to servers or routers that do not run OSPF it is advised to configure them as passive IP interfaces. Use the command `ip ospf passive-interface` to accomplish that.

Configuration example for internal VMS links:

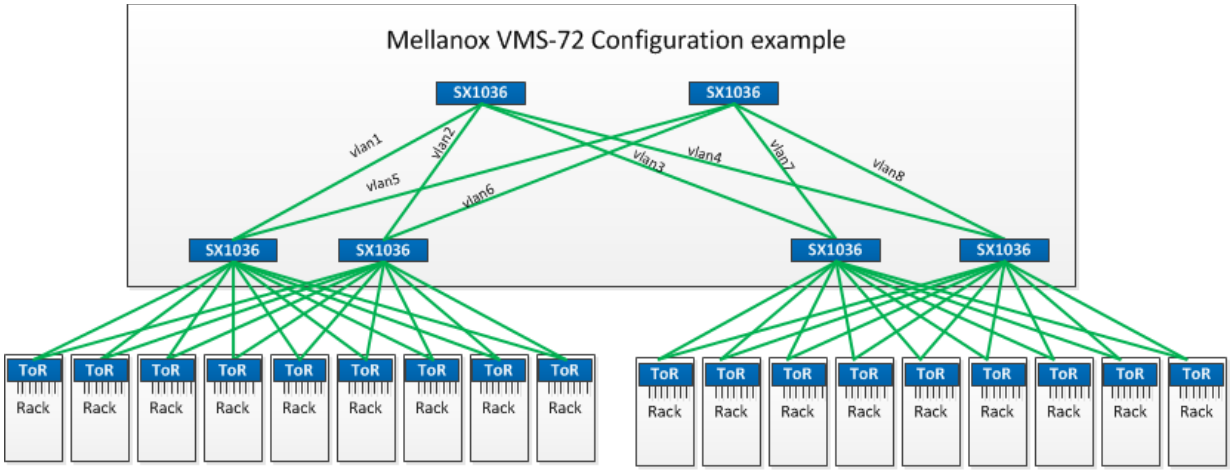
```
switch (config)# protocol ospf
switch (config)# router ospf
switch (config)# interface vlan 10 // Vlan 10 is used between a leaf and a spine
switch (config interface vlan 10)# ip ospf area 0
switch (config interface vlan 10)# ip ospf network point-to-point
switch (config interface vlan 10)# exit
switch (config)# interface vlan 20 // Vlan 20 is used between a leaf and a server
switch (config interface vlan 20)# ip ospf area 0
switch (config interface vlan 20)# ip ospf passive-interface
```

Area 0 is the backbone area.

Mellanox® Technologies Confidential

3.3 Configuration Example

Figure 5: VMS-72 Configuration Example



The following steps provide an initial configuration flow necessary on each VMS switch.

➢ ***To configure the VMS:***

Step 1. Ensure the prerequisites are met. Refer to Section 3.1, “Prerequisites,” on page 11.

Step 2. Disable spanning-tree using the command `no spanning-tree`.

Step 3. (Recommended) Enable LLDP. use the command `lldp`.

Step 4. (Optional) For each IP interface, configure the link as a LAG. Refer to Section 3.2.2, “Creating a Static LAG,” on page 12.

Step 5. Set a VLAN for each L2 interface. Refer to Section 3.2.3, “Setting a VLAN Association to L2 interface,” on page 13.

Step 6. Create a VLAN interface (with the same VLAN used in the previous step) for each IP interface. Refer to Section 3.2.1, “Creating a VLAN Interface,” on page 11.

Step 7. Enable OSPF on the ToR. Refer to Section 3.2.4, “Basic OSPF Configuration,” on page 13.

Mellanox® Technologies Confidential

3.3.1 Spine-1

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 1
  exit
interface ethernet 1/1 switchport access vlan 1
interface vlan 1
  ip address 11.11.1.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 2
  exit
interface ethernet 1/2 switchport access vlan 2
interface vlan 2
  ip address 11.11.2.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 3
  exit
interface ethernet 1/3 switchport access vlan 3
interface vlan 3
  ip address 11.11.3.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 4
  exit
interface ethernet 1/4 switchport access vlan 4
interface vlan 4
  ip address 11.11.4.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```


Mellanox® Technologies Confidential

3.3.2 Spine-2

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 5
  exit
interface ethernet 1/1 switchport access vlan 5
interface vlan 5
  ip address 11.11.5.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 6
  exit
interface ethernet 1/2 switchport access vlan 6
interface vlan 6
  ip address 11.11.6.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 7
  exit
interface ethernet 1/3 switchport access vlan 7
interface vlan 7
  ip address 11.11.7.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 8
  exit
interface ethernet 1/4 switchport access vlan 8
interface vlan 8
  ip address 11.11.8.2 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```


Mellanox® Technologies Confidential

3.3.3 Leaf-1

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 1
  exit
interface ethernet 1/1 switchport access vlan 1
interface vlan 1
  ip address 11.11.1.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 5
interfac ethernet 1/2 switchport access vlan 5
interface vlan 5
  ip address 11.11.5.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```

3.3.4 Leaf-2

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 2
  exit
interface ethernet 1/1 switchport access vlan 2
interface vlan 2
  ip address 11.11.2.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 6
interfac ethernet 1/2 switchport access vlan 6
interface vlan 6
  ip address 11.11.6.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```


Mellanox® Technologies Confidential

3.3.5 Leaf-3

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 3
  exit
interface ethernet 1/1 switchport access vlan 3
interface vlan 3
  ip address 11.11.3.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 7
interfac ethernet 1/2 switchport access vlan 7
interface vlan 7
  ip address 11.11.7.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```

3.3.6 Leaf-4

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
  exit

vlan 4
  exit
interface ethernet 1/1 switchport access vlan 4
interface vlan 4
  ip address 11.11.4.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 8
interfac ethernet 1/2 switchport access vlan 8
interface vlan 8
  ip address 11.11.8.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit
```


Mellanox® Technologies Confidential

4 Advanced Considerations

4.1 Load Balancing

Layer 3 load balancing between leafs and spines is achieved via OSPF and equal cost multi-path (ECMP). When using ECMP, the configuration of a LAG is not mandatory since ECMP already performs the load balancing between the links. Using separate links between a leaf and a spine (without grouping them to a LAG) is likely to be the best option for best performance. However, aggregating those links to a LAG results in easier configuration and VMS management because logically the number of interfaces decreases.

The overall throughput depends on the LAG and ECMP distribution function, and hence may not reach a 100%. To calculate the overall throughput, when using both LAG and ECMP, multiply both of the distribution functions (the one from the LAG and the one from the ECMP). That may cause a decrease in the throughput of the network comparing to a topology that does not use LAG configuration. As network scales expand, however, it becomes more practical to use LAG in order to ease the workload of the tables and the CPU.

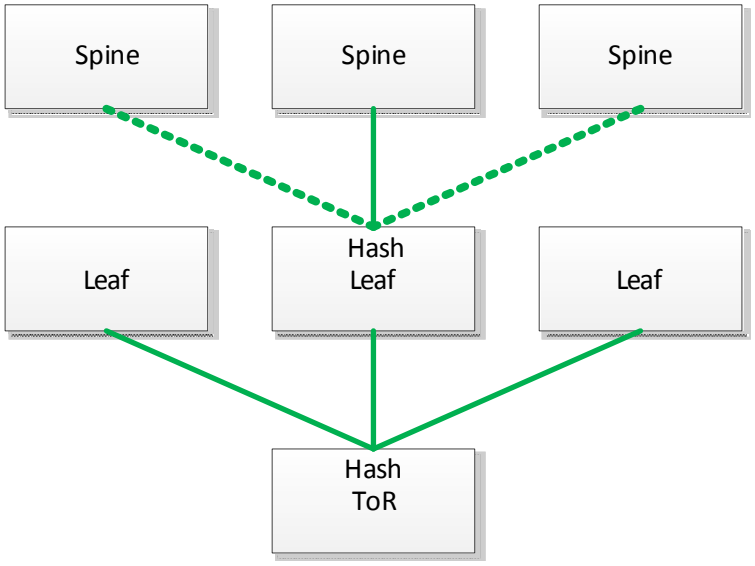
Configuring load balancing to use MAC address hash should be avoided as the VMS is configured in point-to-point L3. In such a case, there will be only one source MAC and one destination MAC and therefore load balancing will have no effect.

For more information about ECMP, please refer to Section 4.1.2, “Equal Cost Multi-Path Routing (ECMP),” on page 20.

4.1.1 Hash Functions

It is advised that LAG and ECMP hash function configuration on the leafs and the spines is different. If the same hash function is used in the ToR and the leaf, all the traffic directed from the leaf to all the spines will arrive only to one of the spines.

Figure 6: Multiple Hush Functions



Mellanox® Technologies Confidential

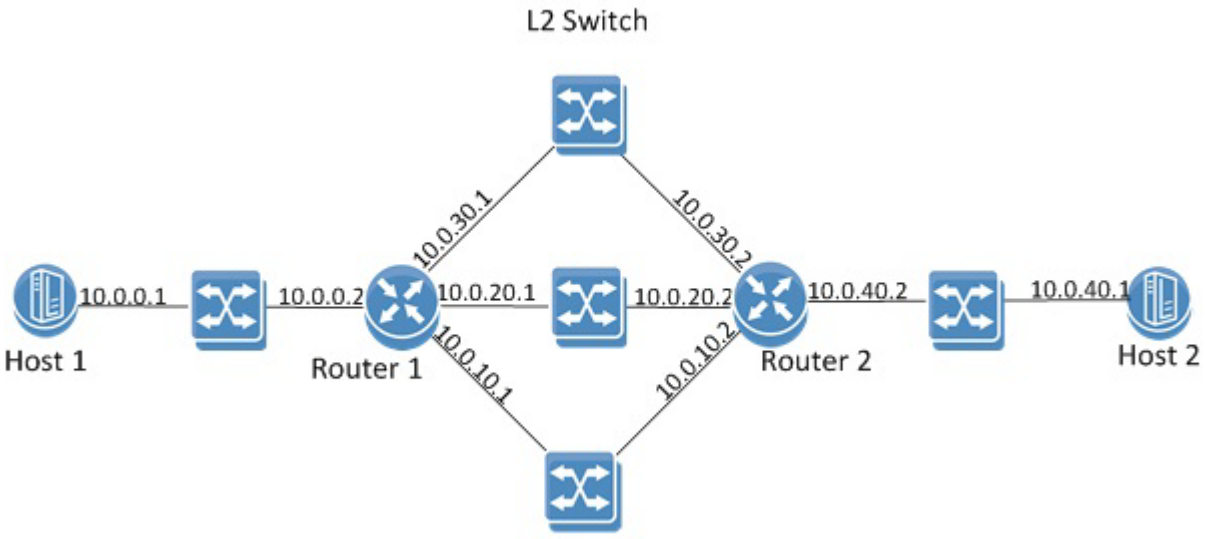
4.1.2 Equal Cost Multi-Path Routing (ECMP)

ECMP is a routing strategy where next-hop packet forwarding to a single destination can occur over multiple paths.

In the figure below, Router 1 and Router 2 can both access each of the networks of the peer router. The routing options of Router 1 for the 10.0.40.1/2 network include the following routes:

- 10.0.10.2
- 10.0.20.2
- 10.0.30.2

Figure 7: ECMP



The load balancing function of ECMP is configured globally on the system.

The ECMP function has a hash algorithm that determines through which route to direct traffic. This hash algorithm can be symmetric or asymmetric. In symmetric hash functions, bidirectional flows between routes follow the same path, while in asymmetric hash functions, bidirectional traffic can follow different paths in both directions.

The following load balancing types are supported:

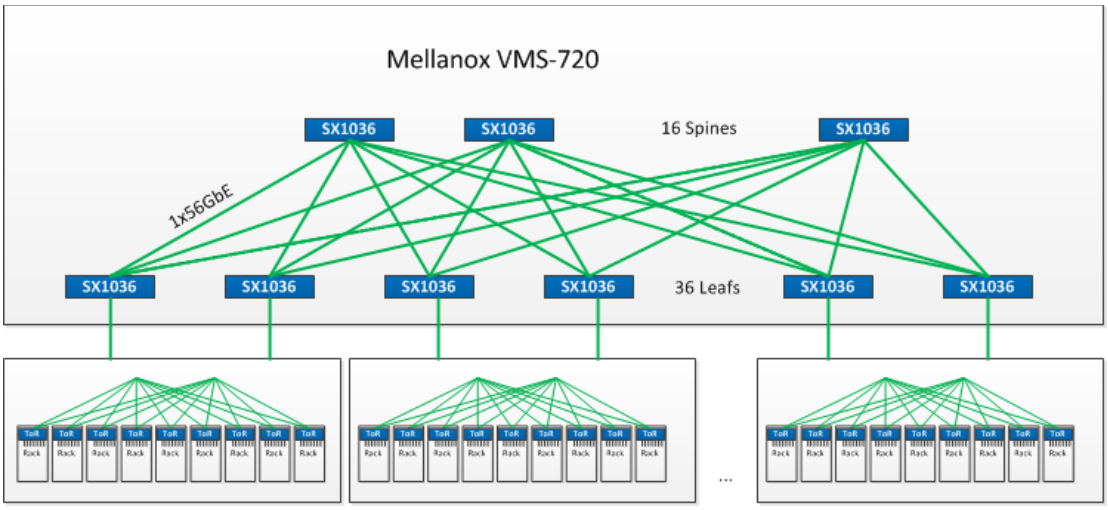
- Source IP & port – source IP (SIP) and source UDP/TCP port: If the packet is not UDP/TCP, only SIP is used for the hash calculation. This is an asymmetric hash function.
- Destination IP & port – destination IP (DIP) and destination UDP/TCP port: If the packet is not UDP/TCP, only DIP is used for the hash calculation. This is an asymmetric hash function.
- Source and destination IP & port – destination and source IP, as well as destination and source UDP/TCP port: If the packet is not UDP/TCP, then SIP/DIP are used for the hash calculation. This is a symmetric hash function.
- Traffic class: Load balancing based on the traffic class assigned to the packet. This is an asymmetric hash function.
- All (default): All the above fields are part of the hash calculations. This is a symmetric hash function.

Mellanox® Technologies Confidential

4.2 56GbE Interface Rate

One of the unique capabilities of the Mellanox® platforms is a proprietary ability to speed up its 40GbE links to run at 56GbE rate. This capability opens up more options for building VMS-720. VMS-720 topology consists of 16 interfaces (out of 36) of the SX1036 switch to be configured on the leaf towards the spines freeing 20 interfaces to be configured on the leaf towards the ToRs. In such a case that the inner ports of the VMS (the links between the leafs and spines) are configured with 56GbE the inner throughput increases. Consequently, more interfaces are free to be used towards the ToRs.

Figure 8: Mellanox VMS-720



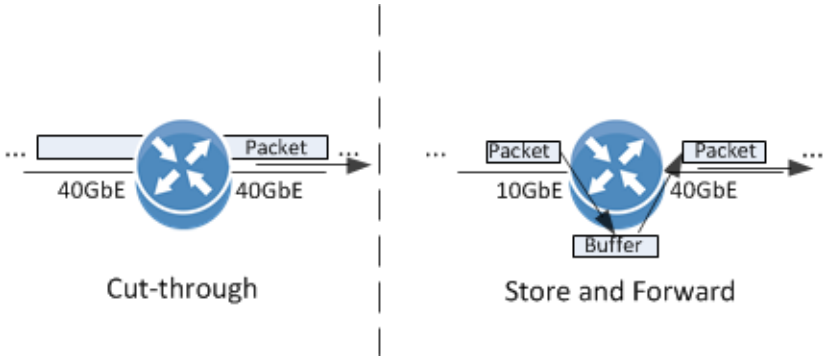
Setting the link speed on Mellanox platforms requires a license. Please refer to *MLNX-OS® User Manual* for detailed information.

4.3 Store and Forward

If low-latency is crucial to a certain network it is recommended to have the leafs and spines configured with similar ingress and egress rate interfaces (e.g. configure them all to be 40GbE). An ingress interface of 10GbE passing traffic to 40GbE interface, for example, causes the router to store the whole packet before sending it which increases latency depending on the size of the packet and the speed of the interface. If the ingress and egress interfaces have the same rate (either 10GbE or 40GbE) the switch operates in cut-through mode. Each packet is parsed and sent even before the whole packet gets into the system. This act reduces overall latency in the network.

Mellanox® Technologies Confidential

Figure 9: Store and Forward



4.4 OSPF

Large VMS setups (such as VMS-648) support many networks that cannot be managed in a single area (for example area 0). Therefore, it is recommended to use several areas to reduce the number of entries in the routing table. It is recommended to configure these areas on the bundled VMS leafs and ToRs while area 0 can be configured on the leafs and the spines.

Additional OSPF considerations may arise per network and solutions can be tailored upon demand.

It is recommended to configure large VMS setups so that every two leafs' external ports are configured under the same area (other than area 0). For example, VMS-648, which has 36 leafs, would have 19 areas in total – 18 areas towards the ToRs, and area 0 configured on the spines and leafs' inner ports. For an illustration, refer to Figure 10, "VMS-72 OSPF Totally Stub Area Example," on page 23.

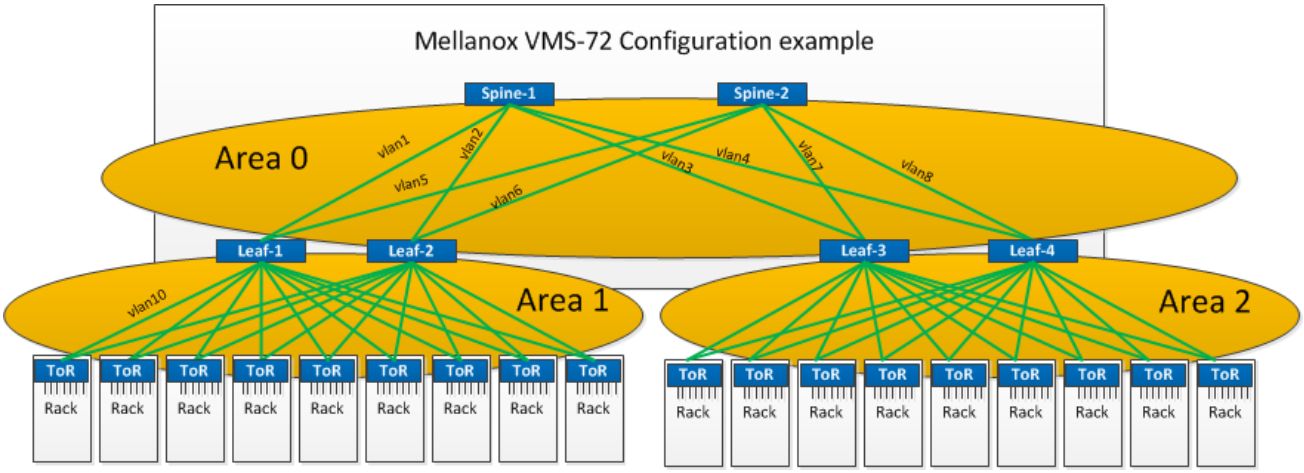
4.4.1 Totally Stub Area OSPF

A stub area is an area which does not receive route advertisements external to the autonomous system and routing from within the area is based entirely on a default route. An Area Boarder Router (ABR) deletes type 4, 5 Link-State Advertisements (LSAs) from internal routers, sends them a default route of 0.0.0.0 and turns itself into a default gateway. This reduces Link-State Database (LSDB) and routing table size for internal routers.

A totally stub area (TSA) is similar to a stub area. However, this area does not allow summary routes in addition to not having external routes. That is, inter-area (IA) routes are not summarized into totally stubby areas. The only way for traffic to get routed outside of the area is a default route which is the only Type-3 LSA advertised into the area. When there is only one route out of the area, fewer routing decisions have to be made by the route processor which lowers system resource utilization. More than one ABR router may be configured in each area. In Figure 10 below, Leaf-1 and Leaf-2 are configured as TSA.

Mellanox® Technologies Confidential

Figure 10: VMS-72 OSPF Totally Stub Area Example



➢ ***To configure OSPF stub areas on the switch:***

Step 1. As a prerequisite, make sure OSPF is running on the switch. Refer to Section 3.2.4, "Basic OSPF Configuration," on page 13.

Step 2. Create a totally stub area. Run the command `area <number> stub no-summary` under `router ospf` configuration mode.

```
switch (config)# router ospf
switch (config router ospf)# area 1 stub no-summary // <- ToRs and leafs
```

Step 3. In all the leafs that act as OSPF ABRs, add a network range area to the totally stub area using the command `area <number> range`.

```
switch (config router ospf)# area 1 range 11.12.0.0 /16 // <- only in ABR (leafs)
```

It is advised to design the network in a way that only one or few summary address will be advertized from the ABRs (leafs) to area 0.

Step 4. Add each IP interface on the switch needed to be attached to the OSPF stub area.

```
switch (config interface vlan 1)# ip ospf area 1
```

Step 5. Follow the same configuration on all leafs and ToRs designed to be attached to area 1.

It is advised to design the network in a way that only 1 (or few) summary addresses will be advertized from the ABRs (leafs) to area 0.

Mellanox® Technologies Confidential

➢ ***To verify the OSPF stub configuration:***

Step 1. To see the configured areas on the switch, run the command `show ip ospf`

```
switch (config) # show ip ospf

...
rea (0.0.0.1) (Active)
Interfaces in this area: 1
  Active Interfaces:  1                          // <-- The interface to the ToR
  Passive Interfaces: 0
SPF Calculation has run 24 times
This area is STUB area
Area Default Cost: 1
Do not send area summary
Number of LSAs: 3, checksum sum 137717

Area (0.0.0.0) (Active)
Interfaces in this area: 2
  Active Interfaces:  2                          // <-- The interfaces to the Spines
  Passive Interfaces: 0
SPF Calculation has run 35 times
This area is Normal area
Number of LSAs: 12, checksum sum 373878

switch (config) #
```

Step 2. Run the command `show ip ospf neighbors` to see the neighbors and their attached areas.

```
switch (config) # show ip ospf neighbors

Neighbor 11.12.10.2, interface address 11.12.10.2
In the area 0.0.0.1  via interface Vlan 10
Neighbor priority is 1, State is FULL
No designated router on this network
No backup designated router on this network
Options 0
Dead timer due in 32

...

switch (config) #
```

Step 3. On the leafs (ABRs), to see the configured summary addresses on the stub area run the command `show ip ospf summary-address`

```
switch (config) # show ip ospf summary-address
OSPF Process ID Default
Network    Mask           Area    Advertise    LSA type  Metric  Tag
-------------------------------------------------------------------
11.12.0.0  255.255.0.0   0.0.0.1  Advertise    Type 3    10      N/A
# switch (config) #
```


Mellanox® Technologies Confidential

The running-config below is an example for leaf1 switch configuration with stub area 1 for IP interface 10 (interface VLAN 10).

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
area 1 stub no-summary
area 1 range 11.12.0.0 /16
  exit

vlan 1
  exit
interface ethernet 1/1 switchport access vlan 1
interface vlan 1
  ip address 11.11.1.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 5
interfac ethernet 1/2 switchport access vlan 5
interface vlan 5
  ip address 11.11.5.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 10
interfac ethernet 1/19 switchport access vlan 10
interface vlan 10
  ip address 11.12.10.2 255.255.255.0
  ip ospf area 1
  ip ospf network point-to-point
  exit
```


Mellanox® Technologies Confidential

The running-config below is an example for ToR1 switch configuration with stub area 1 for IP interface 10 (interface VLAN 10).

```
lldp
no spanning-tree
ip routing
protocol ospf
router ospf
area 1 stub no-summary
  exit

vlan 1
  exit
interface ethernet 1/1 switchport access vlan 1
interface vlan 1
  ip address 11.11.1.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
exit

vlan 5
interfac ethernet 1/2 switchport access vlan 5
interface vlan 5
  ip address 11.11.5.1 255.255.255.0
  ip ospf area 0.0.0.0
  ip ospf network point-to-point
  exit

vlan 10
interfac ethernet 1/1 switchport access vlan 10
interface vlan 10
  ip address 11.12.10.1 255.255.255.0
  ip ospf area 1
  ip ospf network point-to-point
  exit
```

4.5 Automation Tools

Automation of tasks becomes more significant in large scale VMS. IT managers may develop scripts inhouse to help him with automation or use OpenSource tools to do so. Mellanox supports Puppet and Opscode's Chef. Puppet and Chef have been developed to help IT managers build and share mature tools for the automation of different systems.

For additional information on Puppet and Chef visit Mellanox's community website at:

http://community.mellanox.com/docs/DOC-1196

4.6 Network Discovery

Network discovery is an important feature for small and large scale VMS for monitoring and debugging the network. It is advised to turn on LLDP protocol on all the switches (Mellanox and any other 3rd party) and to install the lldpad tool on your servers, and eventually to enable LLDP in your network.

There may be several tools in the market that perform discovery based on LLDP. Mellanox supplies a simple PERL script to do this job.

Mellanox® Technologies Confidential

The script simply uses the SNMP LLDP-MIB to get the remote neighbor of each interface and print them.

In the example below there is an output example of the script that supplies per interface the speed and the remote system or server connected plus the remote interface..

```
# ./lldp_script
 - Reading Devices  from file: list.txt

Discovering topology for the following switches:
1. r-qa-sit-sx01
2. r-qa-sit-sx02
3. r-qa-sit-sx03
4. r-qa-sit-sx04
5. r-qa-sit-sx05
6. r-qa-sit-sx06
7. r-qa-sit-sx104

Topology file created 2013-07-04-09_33_51.etopo

r-qa-sit-sx06
        Eth1/1 40000 -> r-qa-sit-sx03 Eth1/2
        Eth1/2 40000 -> r-qa-sit-sx04 Eth1/2

r-qa-sit-sx03
        Eth1/1 40000 -> r-qa-sit-sx05 Eth1/1
        Eth1/10 40000 -> r-qa-sit-sx104 Eth1/1
        Eth1/2 40000 -> r-qa-sit-sx06 Eth1/1
        Eth1/20 40000 -> r-qa-sit-sx01 Eth1/49

r-qa-sit-sx04
        Eth1/1 40000 -> r-qa-sit-sx05 Eth1/2
        Eth1/2 40000 -> r-qa-sit-sx06 Eth1/2
        Eth1/20 40000 -> r-qa-sit-sx02 Eth1/1

r-qa-sit-sx02
        Eth1/1 40000 -> r-qa-sit-sx04 Eth1/20
        Eth1/20 40000 -> reg-r-vrt-002

r-qa-sit-sx01
        Eth1/49 40000 -> r-qa-sit-sx03 Eth1/20
        Eth1/60 40000 -> reg-r-vrt-001

r-qa-sit-sx104
        Eth1/1 40000 -> r-qa-sit-sx03 Eth1/10

r-qa-sit-sx05
        Eth1/1 40000 -> r-qa-sit-sx03 Eth1/1
        Eth1/2 40000 -> r-qa-sit-sx04 Eth1/1
[root@reg-r-vrt-001 ~]#
```

For additional information on the discovery script refer to Mellanox's community website at:

http://community.mellanox.com/docs/DOC-1197

Mellanox® Technologies Confidential

Appendix A: Ordering Information

Mellanox offers the switch systems in Table 4. Depending on the topology, one is able to select the most suitable switches.

Table 4 - Supported Switch Systems

Part Number	Description
MSX1036-xxS	36-Port QSFP 40GigE 1U Ethernet switch. http://www.mellanox.com/page/products_dyn?product_family=115
MSX1024-xxS	48-Port SFP+ 10GigE, 12-Port QSFP 40GigE 1U Ethernet switch http://www.mellanox.com/page/products_dyn?product_family=130
MSX1012-xxS	12-Port QSFP 40GigE 1U Ethernet switch. http://www.mellanox.com/page/products_dyn?product_family=163

For setup and installation refer to *MLNX-OS® User Manual,* and *MLNX-OS® Command Reference Guide* found on support.mellanox.com > Software & Drivers > Management Software > MLNX-OS®.

Please refer to the *Mellanox Products Approved Cable Lists* document for the list of supported cables. http://www.mellanox.com/related-docs/user_manuals/Mellanox_approved_cables.pdf

A.1 Licenses

To reach 56GbE rate the following licence should be installed.

Table 5 - MLNX-OS Licenses

OPN	Description
LIC-1036-56GE	56GbE support for SX1036.
LIC-1024-56GE	56GbE support for SX1024.
LIC-1012-56GE	56GbE support for SX1012.

Mellanox® Technologies Confidential

Appendix B: Mellanox VMS – Basic POC

In order to get started with a basic proof of concept (POC) VMS-72 can be considered which requires the hardware delineated in Table 6.

Table 6 - Mellanox VMS Switch Systems

P/N	Description	Quantity
MSX1036-XXX	36-Port QSFP 40GigE 1U Ethernet switch.	4
MC2207130-XXX	Passive copper cable, VPI, up to 56Gb/s, QSFP.	4

Two servers must be connected to the VMS, with or without TOR switches.

Mellanox® Technologies Confidential